Multiple folding pathways of proteins with shallow knots and co-translational folding

Mateusz Chwastyk and Marek Cieplak

Institute of Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warsaw, Poland

September 4, 2015

Abstract

We study the folding process in the shallowly knotted protein MJ0366 within two variants of a structure-based model. We observe that the resulting topological pathways are much richer than identified in previous studies. In addition to the single knot-loop events, we find novel, and dominant, two-loop mechanisms. We demonstrate that folding takes place in a range of temperatures and the conditions of most successful folding are at temperatures which are higher than those required for the fastest folding. We also demonstrate that nascent conditions are more favorable to knotting than off-ribosome folding.

1 Introduction

Sufficiently long polymers, such as DNA, are likely to be entangled [1, 2]. On the other hand, there are very few knotted RNA molecules – in a recent assessment [3] only three cases have been identified, but the corresponding structures are resolved poorly. Occurrence of knotted proteins is in between – several hundreds of knot-containing structure files and a dozen of truly independent structures [4, 5, 6, 7] in the Protein Data Bank (PDB). Knots in proteins can be detected experimentally through stretching [8] since their effective contour length is reduced, as analyzed theoretically in refs. [9, 10, 11, 12]. However, figuring out how knots get formed during the folding process is more challenging. New experimental techniques based on fusion proteins [13, 14] have started to offer clues about the process. Nevertheless computer simulations are expected to offer more detailed insights. In this paper, we analyze pathways of folding in a model shallowly knotted protein and reveal a rich complexity of possible behaviors. We also propose a mechanical model of the ribosome and show that nascent conditions favor knotting but also affect the pathways of folding.

Backbones of proteins do not form closed loops which leads to some ambiguities when deciding about the presence of a knot. Nevertheless, it is often straightforward to identify knot ends by observing knot's unknotting on cutting away sites from the termini [17, 18]. A knot is considered shallow if at least one of its ends is close to a terminus. Otherwise, the knot is considered to be deep.

Deeply knotted proteins such as YibK have been studied experimentally [15, 16]. They are known to fold with difficulty in simulations. A folding process here is considered successful if the native contacts are established *and* the knot is formed properly. The success rate, S, is, at best, 1 - 2 % [19, 20]. We have recently argued [21] that nascent conditions [22, 23, 24, 25, 26] enhance the probability for YibK to become knotted when formed on the ribosome. We have confirmed that the successful folding pathway goes through a slipknot conformation, as suggested in ref. [19] for the off-ribosome situation.

We have shown that the process takes place only in a range of optimal temperatures, and does not require any non-native interactions (if a proper procedure for their selection is adopted).

In this paper, we focus on a protein with the native shallow trefoil knot: MJ0366 from *Methanocaldococcus jannaschii* which is a thermophilic methanogenic archaea. This is the smallest of the known knotted proteins and we shall refer to this protein by its PDB structure code of 2EFV. Folding of 2EFV has been studied theoretically [27, 28] at a fixed temperature (T). The studies involved all-atom simulations. Ref. [27] used a simplifying implicit solvent approach combined with a bias implemented through the dominant reaction pathway method. Out of 32 successful trajectories, 26 involved direct threading (DT), 3 – slipknotting (SK), and 2 – mousetrapping (MT) as mechanisms of knotting. These mechanisms are illustrated in Figure 1. On the other hand, the simulations in ref. [28] took a slipknot conformation as the initial state of the system without exploring other possible pathways: out of 15 40-μs-simulations, 5 resulted in correct folding.

Here, we use two variants of a structure-based coarse-grained model, in which the protein is represented as a chain of beads located at the α-C positions, and consider various T's and much larger statistics of between 100 to 300 trajectories for each T. We show that: a) 2EFV gets to the knotted native state much easier than the deeply knotted proteins, b) 2EFV should fold through a qualitatively richer family of pathways than considered in ref. [27], c) there is a range of optimal T's for successful knotting, d) the T-range corresponding to the fastest folding is shifted downward relative to the T-range of the optimal knotting, e) nascent conditions boost the peak success rate to fold in one of the variants of the model (in the other variant the off-ribosome peak success rate is already 100%), and f) nascent conditions reduce the set of knotting pathways.

All of the knotting mechanisms identified in ref. [27] are topologically single-stage processes in which just one knot-loop is formed. We find that about 40% of our successful trajectories indeed belong to this class. However, the majority of the trajectories involve two stages and two smaller knot-loops. Each of the stages makes use of variants of the DT, and SK events, but we also identify one more - an "embracement" (EM). It is possible that the multiple-loop mechanisms of knotting are also relevant for the homopolymer-like DNA, but have not been identified yet.

2 Structure-based modeling

The details of our approach are described in refs. [29, 30, 31]. The model is Go-like [32] so that the length-related parameters in the potentials are derived from the native structure. The molecular dynamics employed deals only with the α-C atoms. The bonded interactions are described by the harmonic potentials. Non-bonded interactions, or contacts, are assigned to pairs of amino acids by using the overlap criterion in which the heavy atoms in the native conformation are represented by enlarged van der Waals spheres [29, 33]: if at least two such spheres from different residues overlap we declare existence of a native contact. These contacts are described by potentials with the minima at the crystallographically determined distances. The potentials are identical in depth, denoted as ϵ. Non-native contacts are considered repulsive.

In order to test the robustness of our results, we consider two variants of the model: C and A. In model C, the contact potentials are of the Lennard-Jones form and the backbone stiffness is accounted for by the chirality potential [29] which favors the native sense of the local backbone chirality. The value of ϵ has been calibrated by making comparisons to the experimental data on stretching: approximately, ϵ/Å is 110 pN (which also is close to the energy of the O-H-N hydrogen bond of 1.65 kcal/mol). In model A, the contact potentials are of the 10-12 kind and the backbone stiffness is described by the more common bond and dihedral angles with the parameters specified in ref. [34]. Model C does not have the bond angle part of model A [30] and in model A, there are no $i, i+3$ contacts. The simulations are done at various temperatures. For most unknotted proteins, optimal folding takes place around $T = 0.3\ \epsilon/k_B$ in model C (see also ref. [35]) and around $T = 0.6\ \epsilon/k_B$ in model A (k_B is the Boltzmann constant; the stiffness parameters depend on ϵ). Both characteristic values of T should correspond to a vicinity of the room T in the respective models.

We use the Langevin thermostat with substantial damping. The time unit of the simulations, τ, is effectively of order 1 ns as the motion of the atoms is dominated by diffusion instead of being ballistic. Folding is usually declared when all native contacts are established for the first time (the distance between two α-C in a contact is smaller than the native distance multiplied by 1.5). For knotted proteins, however, this condition does not necessarily signify that the correct native knot has been formed. The situation in which there is no knot but all contacts are established is referred to as misfolding.

3 Folding of 2EFV

Protein 2EFV comprises 87 residues but the atomic coordinates of the first five of them are not provided in the structure file. In ref. [28], the authors extend the C-terminus (i.e. not where the residues are missing) by a 5-piece helical segment to enhance the definition of the starting slipknot conformation. However, we find this procedure to deteriorate folding properties so we show only the results obtained without such an extension. The secondary of 2EFV consists of 4 helices (23-32, 41-49, 62-71, 74-86) 2 3-10 helices (33-35, 72-73), and 2 β-strand (12-17, 54-59). The knot ends in 2EFV are located at 11 and 73.

Figure 2 summarises the properties of 2EFV in model C. The inset in the lower panel refers to the equilibrium quantities. It shows the probability, P_0, of all native contacts being *simultaneously* established as a function of T. Similar to the lattice models of proteins [36] (see also an exact analysis [37]), one may define T_f as a temperature at which P_0 crosses through $\frac{1}{2}$ – it is 0.22 ϵ/k_B in this model. The T at which the fraction of the established native contacts, Q, crosses through $\frac{1}{2}$ is higher, 0.75 ϵ/k_B, as it signifies the on-cooling onset of globular conformations. This T will be denoted as T_Q.

The top panel of Figure 2 shows the percentage-wise success, S, of reaching the properly knotted folded conformation and the corresponding median folding time t_f as a function of T. The median times have been determined only within the subset of trajectories that resulted in folding. The starting conformations are nearly fully extended. Folding is seen to take place fast in the T-range between 0.2 and 0.55 ϵ/k_B. However, the majority of the trajectories result in knotting only between 0.45 and 0.5 ϵ/k_B where P_0 is 0. Thus the peak success rates involve folding times that are longer than optimal. At $T = 0.3\ \epsilon/k_B$, S is about 5% which is still better than the peak success rates reported for the deeply knotted proteins. The bottom panel shows the S corresponding to the misfolding events. In the T-range corresponding to the optimal t_f's most of the trajectories result in misfolding. S for correct folding and S for misfolding do not add up to 100% as a portion of the trajectories does not establish all native contacts within a cutoff time of 1 000 000 τ.

Figure 3 summarises the properties of 2EFV in model A. The characteristic temperatures that relate to the kinetics move upward by some 0.3 ϵ/k_B and T_f shifts to 0.4 ϵ/k_B while T_Q shifts to 1.0 ϵ/k_B. Even though the peak S for folding is achieved just at the upper edge of the kinetic optimality (at $T = 0.9\epsilon/k_B$), S is larger than 50 % in the whole range of the optimal kinetics (from 0.75 to 0.9 ϵ/k_B. The peak value of S for misfolding (68 %) is at $T = 0.6\epsilon/k_B$ and it disappears at 0.9 ϵ/k_B in a gradual way.

Interestingly, we find that attachment of the 5-, 10-, 15-, and 20-residue N-terminal extensions destroys the proper folding completely – we would expect that all-atom folding with the extensions should be even harder because of the many more degrees of freedom that need to cooperate. However ref. [28] states the opposite. The extensions that we have considered are either random alanine segments or a 5-residue helical extension (88-SER, 89-ALA, 90-ASN, 91-LEU, 92-LEU). The misfolding events, on the other hand, are observed to be frequent.

We observe that shallowly knotted proteins fold and knot much easier than the deeply knotted ones – but what are the mechanisms involved? Figure 4 shows that the proper pathways fall into two classes: with a single knot-loop (the left hand side of the figure) formed from the segment between sites 16–78 or with two knot-loops (the right hand side for the figure) between sites 16–53 (shown in red) and 53–78 (in blue). The two classes arise in both models. In model C, the two-loop mechanisms occur with the T-averaged probability of 58% and in model A – 63%. In the single-loop class, the topological events involve the C-terminal parts. In the DT mechanisms, the C-terminus threads through the knot-loop (45 % in model C). In the SK mechanism, the C-terminal slipknot slides through the knot-loop (the mechanism invoked for the deeply knotted proteins; 45 %). In the MT mechanism, the knot-loop moves to envelop the C-terminus (10 %). A typical time scale in which the first loop forms (also in the single-loop mechanism) is around 4 000 τ in both models. It takes much longer to form the second loop – typically 100 000 τ more. At high T's (like $T = 1.0\ \epsilon/k_B$) in model A, it takes of order 500 000 τ to form a globular state followed a sudden single-loop knotting.

In the two-loop class of pathways, two smaller knot-loops are formed and the events involve both termini. Typically, the topological transformations start at the N-terminus and, at stage A, split into three pathways. They correspond to the mechanisms of DT (6 %), SK (38 %) and EM (56 %). The latter is one in which segment 53-78 "embraces" the mostly idle N-terminal part and forms the knot-loop around it. The next stage, denoted as B, knotting is completed by engaging the C-terminal segment. Here, the pathways split into four mechanisms: SK (44%), MT (11%), DT (39%), and EM (6%). There are some events in which stages A and B are interchanged. An example of a situation in which one EM follows another EM mechanism is shown in Figure 5.

In ref. [27] the knotting mechanisms have been apparently classified based on the events that shortly precede the appearance of knots without identifying the one- and two-loop pathways. If we cumulate the four IIB events with the I events and take the weighted average between the two classes of pathways, we get about 44% in SK, 41% in DT, 11% in MT, and 4% in EM modes which does not agree with about 81% in DT obtained in ref. [27]. This could be due to either the role of the side groups, or the choice of just one T which may not be optimal, or the differences in the statistics.

We now consider a start from the slipknotted conformations without the extensions. If we start from SK - IIB, we get a 96% success rate at $T = 0.35\epsilon/k_B$ and 98% at $T = 0.45\ \epsilon/k_B$. On the other hand, if we starts from SK - IIA then the success rates are correspondingly 34% and 96%. Finally, if we start from SK - I, similar to ref. [28], we get 97 % at $T = 0.45\ \epsilon/k_B$. Our success rate is three times higher than in the all-atom simulations, which may reflect the Go-like character of our model and the lack of the water molecules. However Noel *et al.* [28] extend the protein at the C-terminus which we find to deteriorate folding.

4 Folding of 2EFV under nascent conditions

The percentage-wise success, S, of reaching the properly knotted folded conformation in model C increases substantially when simulating the process in the co-translational way. In our previous model of on-ribosome folding [21] we have focused on the most essential aspects: the excluded volume provided by the ribosome and the related reduction in the conformational entropy are captured by representing the ribosome as an infinite plate which spawns a protein residue by residue at one fixed location. We take the plate to generate a laterally uniform potential of the form $\frac{3\sqrt{3}}{2}\ \epsilon\ (\frac{\sigma_0}{z})^9$, where z denotes the distance away from the plate and $\sigma_0 = 4 \times 2^{-1/6}$. The proteins are synthesized from the N terminus to the C terminus. The time interval between the emergence of two successive α-C atoms, t_w, is taken as 5000 τ since larger values lead to saturation in S. Once the backbone is formed fully, the protein is released and then evolved up to a cutoff time of 1 000 000 τ. Our model is illustrated by Figure 6 which shows the protein toward the end of the process in which it is being born.

Figure 7 shows that the nascent conditions enhance the probability of establishing the native contacts and the peak success rate of knot formation. They also change the look of the dependence of S for knotting on T. For the deeply knotted protein the enhancement in S made a qualitative difference as it enabled folding [21]. For the shallowly knotted 2EFV, the kinetic improvement is minor – by 5 percentage points from the peak value of S=76 %. However, we observe a significant shift in the occurrence of the topological events: the single-loop pathways disappear and the role of the SK events, so crucial for the deeply knotted proteins, gets diminished substantially. Specifically, in step IIA shown in the right hand side of Figure 4 the DT events disappear, the SK events are reduced to 5% and the EM mechanism is operational in 95% of the successful trajectories. In stage IIB, MT and DT are enhanced (to 20 and 55% respectively) whereas SK gets diminished (25%) and EM disappears. The nascent conditions favor loops that first form at the N-terminus.

5 Conclusions

Our results for shallowly knotted 2EFV indicate easy and correct folding with the topological optimality shifted upward in T relative to the kinetic optimality. We have identified novel two-loop mechanisms of knotting and showed that the topological mechanisms are much richer than discussed before. Their variants may also be operational in homopolymers. In our studies of YibK, we have found that extending the overlap-based contact map by extra contacts identified by the CSU server [38] was vital for the success of on-ribosome folding. However, for 2EFV adding the extra contacts is found not to affect any of the results discussed here. On-ribosome folding is more efficient than off-ribosome and is also more selective in its mechanisms. It would be interesting to consider a realistic variant of co-translational folding – one that takes into account confinement. This is because the ribosome spawns proteins into molecularly sculpted chambers [39], which parallels the physics of encapsulation within GroEL-GroES chaperonins considered by Lim and Jackson [14] in the context of deeply knotted proteins.

Acknowledgments We appreciate useful correspondence with C. Micheletti and J. Trylska. This work has been supported by the National Science Center in Poland under the aegis of the EU Joint Programme in Neurodegenerative Diseases (JPND). The local computer resources were financed by the European Regional Development Fund under the Operational Programme Innovative Economy NanoFun POIG.02.02.00-00-025/09.

References

[1] J. M. Arsuaga, S. Vazquez, S. Trigueros, D. W. Sumners, J. Roca, Proc. Natl. Acad. Sci. USA **99**, 5373-5377 (2002).

[2] D. Marenduzzo, C. Micheletti, E. Orlandini, Physics World 30-34 April, (2013).

[3] C. Micheletti, M. Di Stefano, H. Orland, Proc. Natl. Acad. Sci. USA **112**, 2052-2057 (2015).

[4] P. Virnau, L.A. Mirny, M. Kardar, Plos. Comp. Biol. **2**, e122 (2006).

[5] P. Virnau, A. Mallam, S. Jackson, J. Phys. Cond. Mat. **23**, 033101 (2011).

[6] J. I. Sułkowska, E. J. Rawdon, K. C. Millet, J. N. Onuchic, A. Stasiak, Proc. Natl. Acad. Sci. USA **109**, E1715-E1723 (2012).

[7] M. Jamroz, W. Niemyska, E. J. Rawdon, A. Stasiak, K. C. Millet, P. Sułkowski, J. I. Sułkowska, Nucl. Acid. Res. **43**, D306-314 (2015).

[8] D. M. Anstrom, E. Mey, J. Dziubiella, M. Rief, K. T. Forest, Biophys. J. **96**, 1508-1514 (2009).

[9] J.I. Sułkowska, P. Sułkowski, P. Szymczak, M. Cieplak, Phys. Rev. Lett. **100**, 058106 (2008).

[10] J.I. Sułkowska, P. Sułkowski, P. Szymczak, M. Cieplak, J. Am. Chem. Soc. **132**, 13954-13956 (2010).

[11] J. Dziubiella, Biophys. J. **96**, 831-839 (2009).

[12] M. Chwastyk, M. Cieplak, Israel Journal of Chemistry **54**, 1241-1249 (2014).

[13] A. L. Mallam, S. C. Onucha, J. G. Grossman, S. E. Jackson, Mol. Cell. **30**, 642-648 (2008).

[14] N. C. H. Lim, S. E. Jackson, J. Mol. Biol. **427**, 248-258 (2015).

[15] A. L. Mallam, S. E. Jackson, Nat. Chem. Biol. **8**, 147-153 (2012).

[16] A. L. Mallam, J. M. Rogers, S. E. Jackson, Proc. Natl. Acad. Sci. USA **107**, 8189-8194 (2010).

[17] K. Koniaris, M. Muthukumar, Phys. Rev. Lett. **66**, 2211-2214 (1991).

[18] W. Taylor, Nature **406**, 916-919 (2000).

[19] J. I. Sułkowska, P. Sułkowski, J. N. Onuchic, Proc. Natl. Acad. Sci. USA **106**, 3119-3124 (2009).

[20] W. Li, T. Terakawa, W. Wang, S. Takada, Proc. Natl. Acad. Sci. USA **109**, 17789-17794 (2012).

[21] M. Chwastyk, M. Cieplak, *Cotranslational folding of deeply knotted proteins* J. Phys. Cond. Mat. **27**, 354105 (2015).

[22] L. D. Cabrita, C. M. Dobson, J. Christodoulou, Curr. Op. Struct. Biol. **20**, 33-45 (2010).

[23] C. M. Kaiser, D. H. Goldman, J. D. Chodera, I. Tinoco Jr., C. Bustamante, Science **334**, 1723-1727 (2011).

[24] S. Melnikov, A. Ben-Shem, N. Garreau de Loubresse, L. Jenner, G. Yusupova, M. Yusupov, Nat. Struct. Mol. Biol. **19**, 560-567 (2012).

[25] N. Garreau de Loubresse, I. Prokhorova, W. Holtkamp, M. V. Rodnina, G. Yusupova, M. Yusupov, Nature **513**, 517-522 (2014).

[26] J. D. Puglisi, Science **348**, 399-400 (2015).

[27] S. a Beccara, T. Skrbic, R. Covino, C. Micheletti, P. Faccioli, PLOS Comp. Biol. **9**, e1003002 (2013).

[28] J. K. Noel, J. N. Onuchci, J. I. Sułkowska, Phys. Chem. Lett. **4**, 3570-3573 (2013).

[29] J.I. Sułkowska and M. Cieplak, J. Phys.: Cond. Mat. **19**, 283201 (2007).

[30] J. I. Sułkowska, M. Cieplak, Biophys. J. **95**, 3174-3191 (2008).

[31] M. Sikora, J.I. Sułkowska, M. Cieplak, PLOS Comp. Biol. **5**, e1000547 (2009).

[32] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183-210 (1983).

[33] J. Tsai, R. Taylor, C. Chothia, M. Gerstein, J. Mol. Biol. **290**, 253-266 (1999).

[34] C. Clementi, H. Nymeyer, J.N. Onuchic, J. Mol. Biol. **298**, 937-953 (2000).

[35] M. Cieplak, T. X. Hoang, Biophys. J. **84**, 475-488 (2003).

[36] N. D. Socci, J. N. Onuchic, J. Chem. Phys. **101**, 1519-1528 (1994).

[37] M. Cieplak and J. R. Banavar, Phys. Rev. E. Rapid Comm. **88**, 040702(R) (2013).

[38] V. Sobolev, A. Sorokine, J. Prilusky, E.E. Abola, M. Edelman, Bioinformatics **15**, 327-332 (1999).

[39] A. H. Elcock, PLOS Comp. Biol. **2**, e98 (2006).

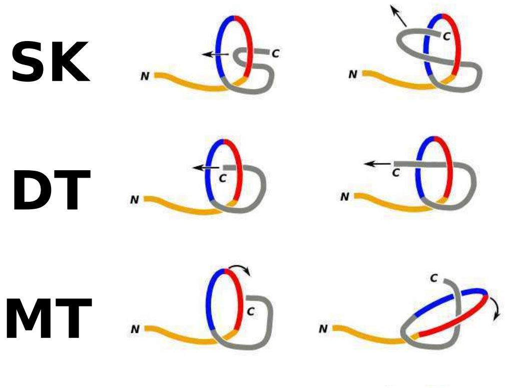

Figure 1: The folding mechanisms in 2EFV found in ref. [27]. SK denotes folding through slipknotting, DT – through direct threading, and MT – through the mousetrap-like mechanism. Each process is illustrated by showing two subsequent stages. SK involves sliding of a slipknot through the knot-loop. In DT, the terminus threads through the knot-loop. MT is similar to DT but the knot-loop makes the dominant movement instead of the terminal site. The orange segment extends from the terminal N to site 16, the red segment – from 17 to 53, the blue segment – from 54 to 78, and the gray segment is the remaining C-terminal piece.

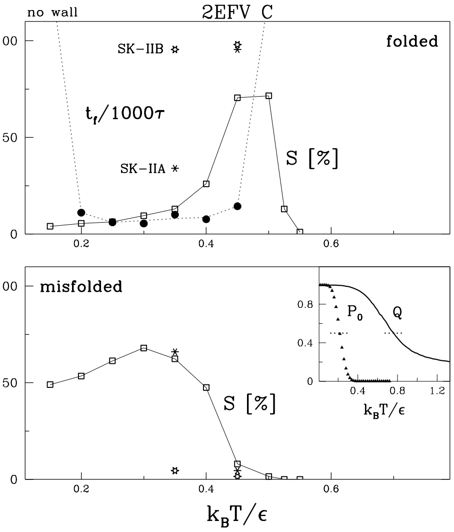

Figure 2: Properties of 2EFV in model C. The inset in the lower panels shows P_0 and Q as a function of T. The main panels characterize the kinetic quantities: the upper panels are for correct folding and the lower panels – for misfolding. The open squares correspond to S – the success rate. The solid circles correspond to the median folding times. These data points are obtained by starting from extended conformations. The data points denoted by the hexagons and asterisks correspond the the slipknotted conformations shown in Figure 1.

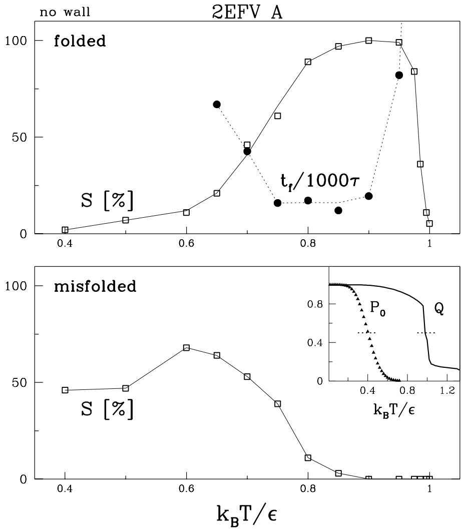

Figure 3: Similar to Figure 2 but for model A.

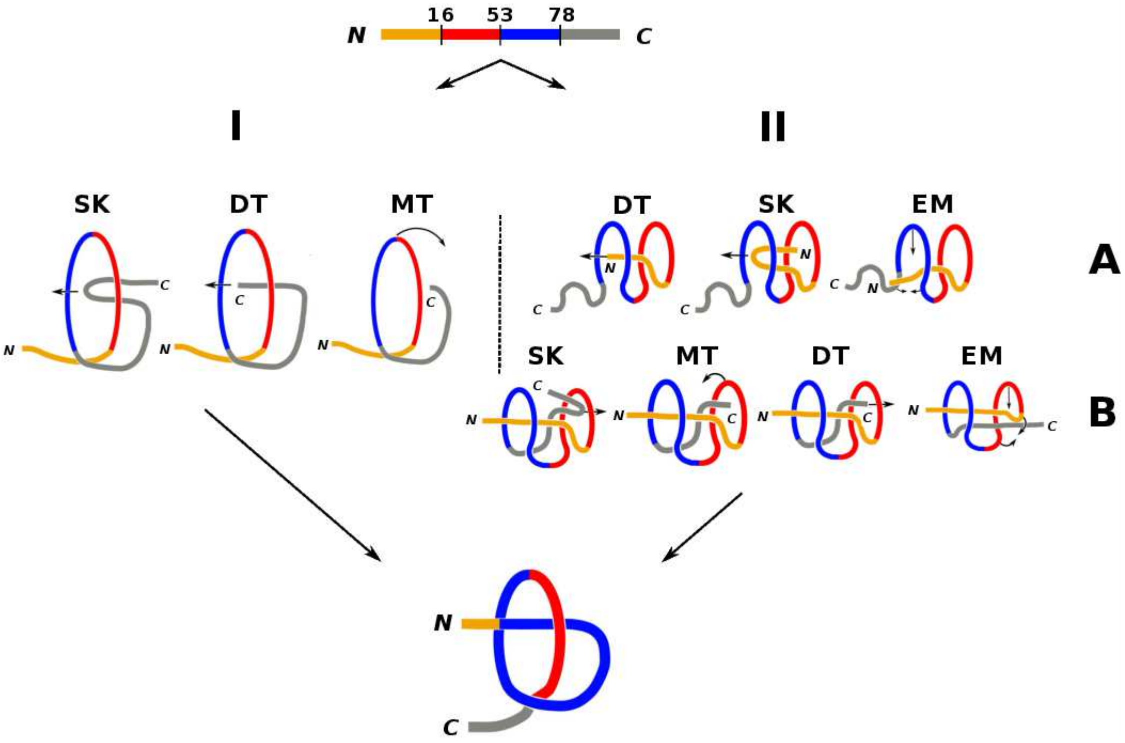

Figure 4: Stages of folding and knotting process of protein 2EFV. The top panel shows the extended conformation which is divided into 4 sequential segments: the N-terminal part is in yellow, C-teminal part is in gray, the first knot-loop is in red, and the second knot loop in blue. The bottom panel gives a schematic representation of the final native state. Block I, on the left, shows the three single-loop mechanisms of knotting: DT – direct threading, SK – slipknotting, MT – mousetrapping. Block II, on the right, shows two stages, A and B, of the two-loop mechanisms of knotting. In addition to DT, SK, and MT, they also involve EM – embracement. The smaller loops have radii between 8 and 10 Å.

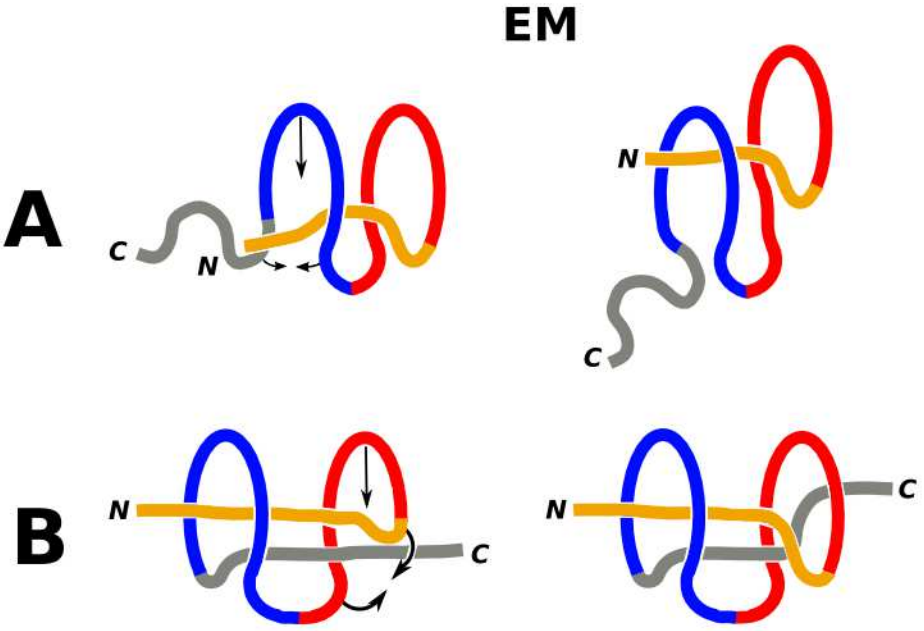

Figure 5: An example of a two-loop mechanisms of folding combined with knotting. Here the embracement of the N-terminus (the top panel) is followed by a similar event at the C-terminus (the bottom panel). The panels on the left show the protein's conformation at the beginning of the knotting at the particular stage. The right panels show the resulting state.

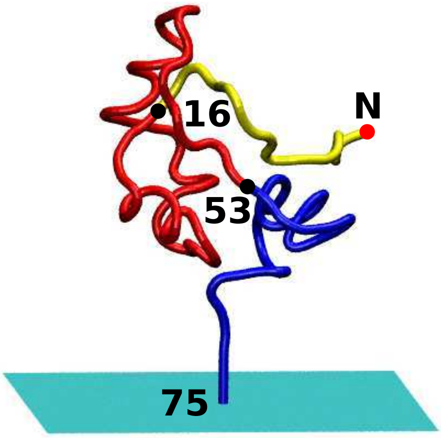

Figure 6: Nascent protein 2EFV emerging from the model ribosome. At the stage shown, the segment 75 – C is not yet born. The colors of the segments in the backbone approximate those of Figure 4.

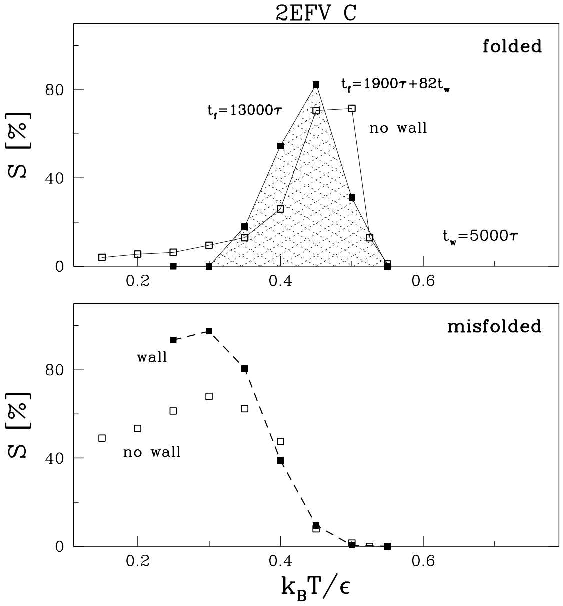

Figure 7: The upper panel shows the success rates for the on-ribosome (the solid squares) and off-ribosome (the open squares) folding as a function of T. The median folding times, t_f, are written for $T = 0.4\ \epsilon/k_B$ (13000 τ is for the off-ribosome folding). The lower panel is similar but shows the data for misfolding.